Rectangular and Nonrectangualr Data

Asad Raza Virk

2024-09-18

Rectangular Data

- The typical frame of reference for an analysis in data science is a rectangular data object, like a spreadsheet or database table
- Rectangular data is the general term for a two-dimensional matrix with rows indicating records (cases) and columns indicating features (variables)
- Data frame is the specific format in R and Python.
- The data doesn't always start in this form: unstructured data (e.g., text) must be processed and manipulated so that it can be represented as a set of features in the rectangular data
- Data in relational databases must be extracted and put into a single table for most data analysis and modeling tasks.

Key Terms for Rectangular Data

1. Data frame Rectangular data (like a spreadsheet) is the basic data structure for statistical and machine learning models.
2. Feature A column within a table is commonly referred to as a feature. Synonyms: attribute, input, predictor, variable
3. Outcome Many data science projects involve predicting an outcome—often a yes/no out- come. The features are sometimes used to predict the outcome in an experiment or a study. Synonyms: dependent variable, response, target, output
4. Records A row within a table is commonly referred to as a record. Synonyms: case, example, instance, observation, pattern, sample

Data Frames and Indexes

Traditional database tables have one or more columns designated as an index, essentially a row number. This can vastly improve the efficiency of certain database queries.

In Python, with the pandas library, the basic rectangular data structure is a DataFrame object. By default, an automatic integer index is created for a DataFrame based on the order of the rows.

In pandas, it is also possible to set multilevel/hierarchical indexes to improve the efficiency of certain operations.

> Terminology Diferences Terminology for rectangular data can be confusing. Statisticians and data scientists use different terms for the same thing. For a statistician, predictor variables are used in a model to predict a response or dependent variable. For a data scientist, features are used to predict a target. One synonym is particularly confusing: computer scientists will use the term sample for a single row; a sample to a statistician means a collection of rows.

Nonrectangular Data Structures

There are other data structures besides rectangular data.

- **Time series data** records successive measurements of the same variable. It is the raw material for statistical forecasting methods, and it is also a key component of the data produced by devices—the Internet of Things.
- **Spatial data** structures, which are used in mapping and location analytics, are more complex and varied than rectangular data structures. In the object representation, the focus of the data is an object (e.g., a house) and its spatial coordinates. The field view, by contrast, focuses on small units of space and the value of a relevant metric (pixel brightness, for example).
- **Graph (or network) data** structures are used to represent physical, social, and abstract relationships. For example, a graph of a social network, such as Facebook or LinkedIn, may represent connections between people on the network. Distribution hubs connected by roads are an example of a physical network. Graph structures are useful for certain types of problems, such as network optimization and recommender systems.

  > Graphs in Statistics In computer science and information technology, the term graph typically refers to a depiction of the connections among entities, and to the underlying data structure. In statistics, graph is used to refer to a variety of plots and visualizations, not just of connections among entities, and the term applies only to the visualization, not to the data structure.